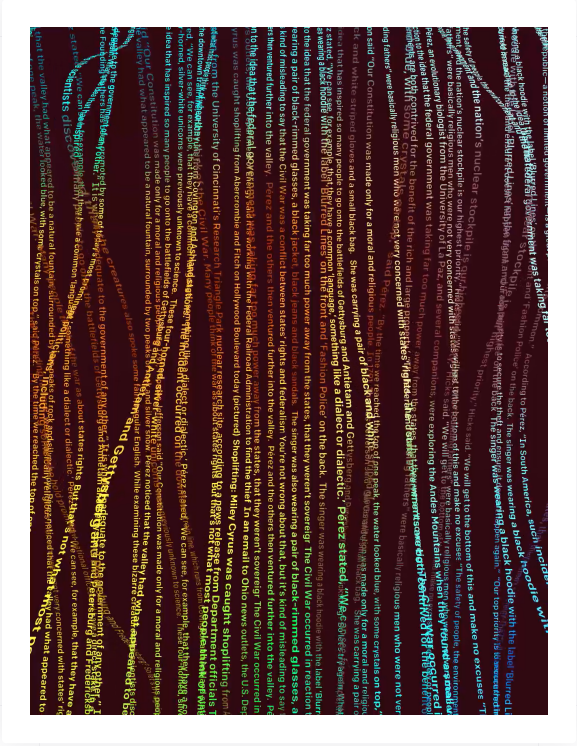


Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

FEBRUARY 14, 2019
24 MINUTE READ

VIEW CODE READ PAPER

Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper.

GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset[1] of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.

GPT-2 displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of unprecedented quality, where we prime the model with an input and have it generate a lengthy continuation. In addition,

We created a new dataset which emphasizes diversity of content, by scraping content from the Internet. In order to preserve document quality, we used only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting (whether educational or funny), leading to higher data quality than other similar datasets, such as CommonCrawl.

GPT-2 outperforms other language models trained on specific domains (like Wikipedia, news, or books) without needing to use these domain-specific training datasets. On language tasks like question answering, reading comprehension, summarization, and translation, GPT-2 begins to learn these tasks from the raw text, using no task-specific training data. While scores on these downstream tasks are far from state-of-the-art, they suggest that the tasks can benefit from unsupervised techniques, given sufficient (unlabeled) data and compute.

Samples

GPT-2 generates synthetic text samples in response to the model being primed with an arbitrary input. The model is chameleon-like—it adapts to the style and content of the conditioning text. This allows the user to generate realistic and coherent continuations about a topic of their choosing, as seen by the following select samples.[2]

Note that while we have hand-chosen these samples, and are thus engaging in some meta-cherry-picking, we believe they are not too unrepresentative of the sampling process. We are simply using top-k truncated sampling, and have yet to explore more advanced methods of sampling (such as beam-search methods).

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

As the above samples show, our model is capable of generating samples from a variety of prompts that feel close to human quality and show coherence over a page or more of text. Nevertheless, we have observed various failure modes, such as repetitive text, world modeling failures (e.g. the model sometimes writes about *fires happening under water*), and unnatural topic switching. Exploring these types of weaknesses of language models is an active area of research in the natural language processing community.

Overall, we find that it takes a few tries to get a good sample, with the number of tries depending on how familiar the model is with the context. When prompted with topics that are highly represented in the data (Brexit, Miley Cyrus, Lord of the Rings, and so on), it seems to be capable of generating reasonable samples about 50% of the time. The opposite is also true: on highly technical or esoteric types of content, the model can perform poorly. Fine-tuning offers the potential for even more detailed control over generated samples—for example, we can fine-tune GPT-2 on the Amazon Reviews dataset and use this to let us write reviews conditioned on things like star rating and category.

These samples have substantial policy implications: large language models are becoming increasingly easy to steer towards scalable, customized, coherent text generation, which in turn could be used in a number of beneficial as well as malicious ways. We'll discuss these implications below in more detail, and outline a publication experiment we are taking in light of such considerations.

Zero-shot

GPT-2 achieves state-of-the-art scores on a variety of domain-specific language modeling tasks. Our model is not trained on any of the data specific to any of these tasks and is only evaluated on them as a final test; this is known as the "zero-shot" setting. GPT-2 outperforms models trained on domain-specific datasets (e.g. Wikipedia, news, books) when evaluated on those same datasets. The following table shows all our state-of-the-art zero-shot results.

(+) means a higher score is better for this domain. (–) means a lower score is better.

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+
LAMBADA	accuracy (+)	**63.24%**	59.23%	95%+
LAMBADA	perplexity (–)	**8.6**	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	**93.30%**	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	**89.05%**	82.3%	92%
Penn Tree Bank	perplexity (–)	**35.76**	46.54	unknown
WikiText-2	perplexity (–)	**18.34**	39.14	unknown
enwik8	bits per character (–)	**0.93**	0.99	unknown
text8	bits per character (–)	**0.98**	1.08	unknown
WikiText-103	perplexity (–)	**17.48**	18.3	unknown

GPT-2 achieves state-of-the-art on Winograd Schema, LAMBADA, and other language modeling tasks.

On other language tasks like question answering, reading comprehension, summarization, and translation, we are able to get surprising results without any fine-tuning of our models, simply by prompting the trained model in the right way (see below for examples of how we do this), though we do still fall short of state-of-the-art for specialized systems.

TASK

Reading Comprehension: answer questions about given passages

DATASET

CoQA

EXAMPLE

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of "one world, one dream". Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the "Journey of Harmony", lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

After being lit at the birthplace of the Olympic Games in Olympia, Greece on March 24, the torch traveled to the Panathinaiko Stadium in Athens, and then to Beijing, arriving on March 31. From Beijing, the torch was following a route passing through six continents. The torch has visited cities along the Silk Road, symbolizing ancient links between China and the rest of the world. The relay also included an ascent with the flame to the top of Mount Everest on the border of Nepal and Tibet, China from the Chinese side, which was closed specially for the event.

Q: What was the theme?
A: "one world, one dream".

Q: What was the length of the race?
A: 137,000 km

Q: Was it larger than previous ones?
A: No

Q: Where did the race begin?
A: Olympia, Greece

Q: Is there anything notable about that place?
A: birthplace of Olympic Games

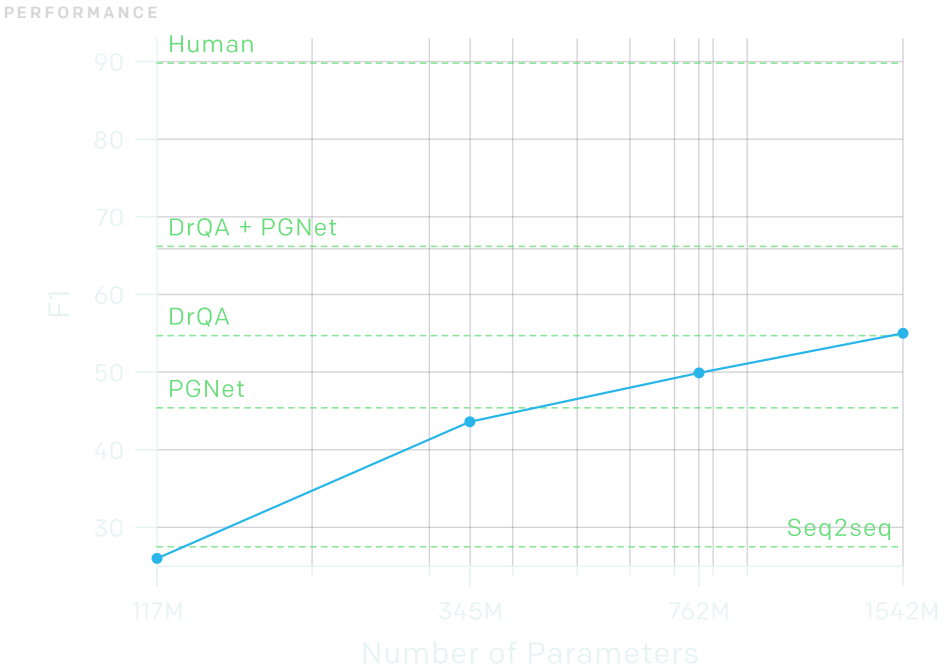
Q: Where did they go after?
A: Athens

Q: How many days was the race?
A: seven

Q: Did they visit any notable landmarks?
A: Panathinaiko Stadium

Q: And did they climb any mountains?
A:

Target answers: *unknown* or *yes*
Model answer: `Everest`



We hypothesize that since these tasks are a subset of general language modeling, we can expect performance to increase further with more compute and data. Others have published similar hypotheses. We also expect fine-tuning to help performance on downstream tasks, though we have yet to do thorough experiments.

Policy Implications

Large, general language models could have significant societal impacts, and also have many near-term applications. We can anticipate how systems like GPT-2 could be used to create:

- AI writing assistants

- More capable dialogue agents
- Unsupervised translation between languages
- Better speech recognition systems

We can also imagine the application of these models for malicious purposes, including the following (or other applications we can't yet anticipate):

- Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content

These findings, combined with earlier results on synthetic imagery, audio, and video, imply that technologies are reducing the cost of generating fake content and waging disinformation campaigns. The public at large will need to become more skeptical of text they find online, just as the "deep fakes" phenomenon calls for more skepticism about images.[3]

Politicians may want to consider introducing penalties for the misuse of such systems, as some have proposed for deep fakes.

Today, malicious actors—some of which are political in nature—have already begun to target the shared online commons, using things like "robotic tools, fake accounts and dedicated teams to troll individuals with hateful commentary or smears that make them afraid to speak, or difficult to be heard or believed". We should consider how research into the generation of synthetic images, videos, audio, and text may further combine to unlock new as-yet-unanticipated capabilities for these actors, and should seek to create better technical and non-technical countermeasures. Furthermore, the underlying technical innovations inherent to these systems are core to fundamental artificial intelligence research, so it is not possible to control research in these domains without slowing down the progress of AI as a whole.

Release Strategy

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research," and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time. This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in

certain research areas. Other disciplines such as biotechnology and cybersecurity have long had active debates about responsible publication in cases with clear misuse potential, and we hope that our experiment will serve as a case study for more nuanced discussions of model and code release decisions in the AI community.

We are aware that some researchers have the technical capacity to reproduce and open source our results. We believe our release strategy limits the initial set of organizations who may choose to do this, and gives the AI community more time to have a discussion about the implications of such systems.

We also think governments should consider expanding or commencing initiatives to more systematically monitor the societal impact and diffusion of AI technologies, and to measure the progression in the capabilities of such systems. If pursued, these efforts could yield a better evidence base for decisions by AI labs and governments regarding publication decisions and AI policy more broadly.

We will further publicly discuss this strategy in six months. If you'd like to discuss large language models and their implications, please email us at: languagequestions@openai.com. And if you're excited about working on cutting-edge language models (and thinking through their policy implications), we're hiring.

GPT-2 Interim Update, May 2019

We're implementing two mechanisms to responsibly publish GPT-2 and hopefully future releases: staged release and partnership-based sharing. We're now releasing a larger 345M version of GPT-2 as a next step in staged release, and are sharing the 762M and 1.5B versions with partners in the AI and security communities who are working to improve societal preparedness for large language models.

Staged Release

Staged release involves the gradual release of a family of models over time. The purpose of our staged release of GPT-2 is to give people time to assess the properties of these models, discuss their societal implications, and evaluate the impacts of release after each stage.

As the next step in our staged release strategy, we are releasing the 345M parameter version of GPT-2. This model features improved performance relative to the 117M version, though falls short of the 1.5B version with respect to the ease of generating coherent text. We have been excited to see so many positive

uses of GPT-2-117M, and hope that 345M will yield still more benefits.

While the misuse risk of 345M is higher than that of 117M, we believe it is substantially lower than that of 1.5B, and we believe that training systems of similar capability to GPT-2-345M is well within the reach of many actors already; this evolving replication landscape has informed our decision-making about what is appropriate to release.

In making our 345M release decision, some of the factors we considered include: the ease of use (by various users) of different model sizes for generating coherent text, the role of humans in the text generation process, the likelihood and timing of future replication and publication by others, evidence of use in the wild and expert-informed inferences about unobservable uses, proofs of concept such as the review generator mentioned in the original blog post, the strength of demand for the models for beneficial purposes, and the input of stakeholders and experts. We remain uncertain about some of these variables and continue to welcome input on how to make appropriate language model publication decisions.

We hope that ongoing research on bias, detection, and misuse will give us the confidence to publish larger models in a timely manner, and at the six month mark we will share a fuller analysis of language models' societal implications and our heuristics for release decisions.

Partnerships

Since releasing this blog post in February, we have had conversations with many external researchers, technology companies, and policymakers about our release strategy and the implications of increasingly large language models. We've also presented or discussed our work at events, including a dinner co-hosted with the Partnership on AI and a presentation to policymakers in Washington DC at the Global Engagement Center.

We are currently forming research partnerships with academic institutions, non-profits, and industry labs focused on increasing societal preparedness for large language models. In particular, we are sharing the 762M and 1.5B parameter versions of GPT-2 to facilitate research on language model output detection, language model bias analysis and mitigation, and analysis of misuse potential. In addition to observing the impacts of language models in the wild, engaging in dialogue with stakeholders, and conducting in-house analysis, these research partnerships will be a key input to our decision-making on larger models. See below for details on how to get involved.

Output Dataset

We're releasing a dataset of GPT-2 outputs from all 4 model sizes, with and

without top-k truncation, as well as a subset of the WebText corpus used to train GPT-2. The output dataset features approximately 250,000 samples per model/hyperparameter pair, which we expect is sufficient to help a wider range of researchers perform quantitative and qualitative analysis on the three topics above. Alongside these datasets, we are including a baseline analysis of some detection-related properties of the models, which we hope others will be able to quickly build on.

Talk to Us

We are interested in collaborating with researchers working on language model output detection, bias, and publication norms, and with organizations potentially affected by large language models: please reach out via our Google Form. Additionally, OpenAI's language, safety, and policy teams will be at ICLR next week, including at the Reproducibility workshop and the OpenAI booth. In particular, we will be discussing this release strategy at the AI for Social Good workshop.

Footnotes

1. We created a new dataset which emphasizes diversity of content, by scraping content from the Internet. In order to preserve document quality, we used only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting (whether educational or funny), leading to higher data quality than other similar datasets, such as CommonCrawl. ↩
2. Note that while we have hand-chosen these samples, and are thus engaging in some meta-cherry-picking, we believe they are not too unrepresentative of the sampling process. We are simply using top-k truncated sampling, and have yet to explore more advanced methods of sampling (such as beam-search methods). ↩
3. Politicians may want to consider introducing penalties for the misuse of such systems, as some have proposed for deep fakes. ↩

Authors

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage & Ilya Sutskever

(ORIGINAL POST)

Miles Brundage, Alec Radford, Jeffrey Wu, Jack Clark, Amanda Askell, David Lansky, Danny Hernandez & David Luan

(INTERIM UPDATE)

Acknowledgments

Thanks to David Luan and Rewon Child for their work on GPT-2.

We also thank the following for feedback on drafts of this post: Greg Brockman, Kai-Fu Lee, Tasha McCauley, Jeffrey Ding, Brian Tse, Allan Dafoe, Rebecca Crootof, Sam Bowman, Ryan Calo, Nick Cammarata and John Schulman.

Editor
Ashley Pilipiszyn

Design
Justin Jay Wang

Cover Artwork
Ben Barry